Transfer learning with prioritized classification and training dataset equalization for medical objects detection

Olga Ostroukhova[1], Konstantin Pogorelov[2,3],
Michael Riegler[3,4], Duc-Tien Dang-Nguyen[5], Pål Halvorsen[3,4]
[1]Research Institute of Multiprocessor Computation Systems n.a. A.V. Kalyaev, Russia
[2]Simula Research Laboratory, Norway [3]University of Oslo, Norway
[4]Simula Metropolitan Center for Digital Engineering, Norway [5]University of Bergen, Norway
olka7lands@gmail.com,konstantin@simula.no,michael@simula.no,ductien.dangnguyen@uib.no,paalh@simula.no

ABSTRACT

This paper presents the method proposed by the organizer team (SIMULA) for *MediaEval 2018 Multimedia for Medicine: the Medico Task*. We utilized the recent transfer-learning-based image classification methodology and focused on how easy it is to implement multi-class image classifiers in general and how to improve the classification performance without deep neural network model redesign. The goal for this was both to provide a baseline for the Medico task and to show the performance of out-of-the-box classifiers for the medical use-case scenario.

1 INTRODUCTION

This paper provides a detailed description of the methods proposed by team SIMULA for MediaEval 2018 Multimedia for Medicine Medico Task [11]. The main goal of the task is to perform medical image classification. The use case scenario is gastrointestinal endoscopies. The 2018-year version of the task is designed as an sixteen classes classification problem. Compared to the 2017-year version which was limited to eight classes [9], the current version of the task comes with several additional challenges such as an imbalanced number of samples in the classes to make it more realistic [8, 9]. In the previous year of the task, participants proposed different methods ranging from simple handcrafted features to deep neural networks [3–6, 10, 12]. For our approach, we propose a convolutional neural network approach (CNN) in combination with transfer learning. To compensate for the imbalanced dataset, we perform prioritized classification and dataset equalization.

2 PROPOSED APPROACH

As the organizer's team for the Medico task, our aim is not achieving the best possible classification performance. Instead, we decided to check how low is the entry threshold to the medical images classification and corresponding lesion detection challenge. To achieve this, and also to provide a baseline for the competing teams, we involved the recent transfer-learning-based image classification methodology and checked how well we are able to (i) easily implement multi-class image classifier and (ii) improve the classification performance without deep neural network model redesign.

Thus, for the basic classification algorithm, we used a CNN architecture and a transfer learning-based classifier, which has been previously introduced for the medical images classification in our previous work [7]. This approach is based on the Inception v3 architecture [13]. To achieve the highest possible performance on the provided limited development set, we used the model pre-trained on the ImageNet dataset [1]. We performed the model retraining using the method described in [2]. We kept all the basic convolutional layers of the network and only retrained the two top fully connected (FC) layers after random initialization of their weights. The FC layers were retrained using the RMSprop [14] optimizer which allows an adaptive learning rate during the training process. We did not used any additional enhancing or pre-processing for the images provided in the datasets. In order to increase the number of training samples, we performed various augmentation operations on the images in the training set. Specifically, we performed horizontal and vertical flipping and a change of brightness in the interval of ±20%.

The initial experimental studies showed that the pre-trained Inception v3 model is able to efficiently extract high-level features from the given medical images, and it is converge quickly during the retraining process with sufficient resulting classification performance (see section 3). However, due to a heavily imbalanced training dataset and despite the used training data augmentation, the detection performance of some classes was not good enough. To solve this issue, we implemented an additional training dataset balancing procedure that performs equalization of the training set by the random duplication of the training samples for the underfilled classes, like *instruments*, *blurry*, etc. This nearly doubled the number of the training samples allowing for better classification performance for the classes with a low number of images provided.

An additional classifier output post-processing step was implemented in order to address the different importance of the different classes as it was stated in the task dataset description [11]. Specifically, we performed the prioritized selection of the resulting output class for each image based of the model's probability output. This was implemented as the selection of the first class with the detection probability higher than a set threshold from the array of classes sorted in order of their importance.

3 RESULTS AND ANALYSIS

For the official task submission creation, two separate models were used, trained on the different datasets. The first model was trained on the training set created from the development set using the described (see section 2) data augmentation procedure. The trained model was used to process the task's test set, and the classification

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

Table 1: Official classification performance evaluation for Detection (D) and Speed (S) runs including ZeroR (ZR), Random (RD) and True (TR) baseline classifiers reporting the following cross-class averaged metrics: True Positive or Hit (TP), True Negative or Correct Rejection (TN), False Positive or False Alarm (FP), False negative or Miss (FN), Recall or Sensitivity or Hit rate or True Positive Rate (REC), Specificity or True Negative Rate (SPE), Precision or Positive Predictive value (PRE), Accuracy (ACC), F1-Score (F1), Matthews correlation coefficient (MCC), Rk statistic or MCC for k different classes (RK), Processing Speed or Frames per Second (FA).

Run	TP	TN	FP	FN	REC	SPE	PRE	ACC	F1	MCC	RK	FPS
D1	474	8122	72	72	0.824	0.991	0.828	0.984	0.815	0.812	**0.854**	43.1
D2	474	8122	72	72	0.823	0.991	0.828	0.984	0.814	0.811	**0.854**	43.0
D3	470	8117	76	76	0.817	0.991	0.819	0.983	0.807	0.803	**0.845**	43.1
D4	440	8087	107	107	0.774	0.987	0.771	0.976	0.756	0.752	**0.786**	43.2
D5	333	7981	213	213	0.664	0.974	0.646	0.951	0.601	0.605	**0.582**	43.0
S1	469	8117	77	77	0.765	0.991	0.729	0.982	0.743	0.737	**0.844**	43.1
S2	469	8117	77	77	0.765	0.991	0.728	0.982	0.743	0.737	**0.844**	43.1
S3	465	8112	82	82	0.758	0.990	0.722	0.981	0.736	0.729	**0.835**	42.9
S4	430	8077	117	117	0.709	0.986	0.677	0.973	0.679	0.674	**0.766**	43.0
S5	313	7960	233	233	0.546	0.971	0.607	0.947	0.504	0.510	**0.544**	43.3
ZR	34	7681	512	512	0.063	0.938	0.004	0.883	0.007	0.0	0.0	-
RD	35	7682	511	511	0.057	0.938	0.064	0.883	0.055	0.001	0.002	-
TR	546	8193	0	0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	-

Table 2: Confusion matrix for the detection run#1 depicted in table 1. The classes are Ulcerative Colitis (A), Esophagitis (B), Normal Z-line (C), Dyed and Lifted Polyps (D), Dyed Resection Margins (E), Out of Patient images (F), Normal Pylorus (G), Stool Inclusions (H), Stool Plenty (I), Blurry Nothing of value (J), Polyps (K), Normal Cecum (L), Colon Clear (M), Retroflex Rectum (N), Retroflex Stomach (O) and Instruments (P).

		Detected class															
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Actual class	A	**459**	2	1	1	5	0	1	0	54	0	13	13	1	7	0	7
	B	2	**388**	77	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	0	145	**451**	0	0	0	4	0	0	0	1	0	0	0	0	0
	D	0	0	0	**406**	81	0	0	0	1	0	4	0	0	0	0	26
	E	0	0	0	115	**462**	0	0	0	0	0	0	1	1	1	0	17
	F	0	0	0	0	1	**2**	0	0	0	0	0	0	0	0	0	0
	G	3	18	27	0	0	0	**548**	0	0	0	2	0	2	1	4	1
	H	10	1	0	5	2	0	0	**498**	98	0	3	1	24	0	0	6
	I	14	0	0	5	1	0	0	0	**1771**	0	5	2	1	3	0	7
	J	2	0	0	0	0	3	0	1	7	**37**	0	0	2	1	0	0
	K	22	1	6	17	2	0	7	1	8	0	**316**	14	1	9	0	64
	L	19	0	0	2	6	0	1	0	16	0	22	**551**	8	3	0	4
	M	3	0	1	1	0	0	0	6	4	0	5	1	**1025**	1	0	6
	N	8	0	0	3	4	0	0	0	3	0	2	1	0	**160**	4	8
	O	0	1	0	0	0	0	0	0	2	0	0	0	0	5	**387**	1
	P	0	0	0	1	0	0	0	0	1	0	1	0	0	1	2	**126**

output was post-processed using the prioritized classification selector with four different probability threshold settings from 0.75 to 0.1 resulting in the runs #2 - #5. For the run #1, we used the max probability selector without class prioritization. The results using the first model were submitted as the speed runs. The second model was trained using the equalized training set, and the same rules for the five runs generation were submitted as the detection run.

The official evaluation results for all the runs are shown in table 1. As one can see, all the runs significantly outperform the ZeroR and Random baselines and show good classification performance. All the runs that utilize the equalized training set have slightly better classification performance. Surprisingly, the introduced prioritized classification method did not result in improved detection performance, not for the original nor for the equalized training sets. With the threshold of 0.75, the classification performance is equal to the non-prioritized runs. It means that the trained classifier is performing as well as it can, and additional re-classification using the class priorities does not make sense for this particular dataset. However, it still can be potentially interesting for bigger datasets or a higher number of classes. The best performing run was the detection run #1 generated using the equalized training set and non-prioritized classifier with the classification performance of 0.854 for Rk statistic (MCC for k different classes). The confusion matrix for this run is depicted in table 2, and the class imbalance and corresponding training and classification challenges can be easily observed. The most challenging class was *Instruments* that is mostly caused by the different shapes, positions and visibilities of the instruments in the images. There also was a number of miss-classification cases for the *Dyed* classes as well as for *Esophagitis* and *Normal Z-line* classes.

With respect to the classification performance in terms of processing speed, the proposed classified can process approximately 43 frames per second on a GPU-enabled consumer-grade personal computer regardless of the enabled or disabled post-processing classes prioritization.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we presented an out-of-the-box solution utilizing a modern pre-trained CNN for the task of medical image classification. The goal was to provide a baseline for the task and to show the performance of basic methods without any deep architecture modification. The best achieved performance measured as Matthew correlation coefficient for k different classes of 0.854 and a speed of 43 frames per second. This is already a quite good result for an out-of-the-box method.

REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* Ieee, 248–255.

[2] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.. In *Proc. of ICML*, Vol. 32. 647–655.

[3] Yang Liu, Zhonglei Gu, and William K Cheung. 2017. HKBU at MediaEval 2017 Medico: Medical multimedia task. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017).*

[4] Syed Sadiq Ali Naqvi, Shees Nadeem, Muhammad Zaid, and Muhammad Atif Tahir. 2017. Ensemble of Texture Features for Finding Abnormalities in the Gastro-Intestinal Tract. *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017).*

[5] Stefan Petscharnig and Klaus Schöffmann. 2018. Learning laparoscopic video shot classification for gynecological surgery. *An International Journal of Multimedia Tools and Applications* 77, 7 (2018), 8061–8079.

[6] Stefan Petscharnig, Klaus Schöffmann, and Mathias Lux. 2017. An Inception-like CNN Architecture for GI Disease and Anatomical Landmark Classification. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017).*

[7] Konstantin Pogorelov, Sigrun Losada Eskeland, Thomas de Lange, Carsten Griwodz, Kristin Ranheim Randel, Håkon Kvale Stensland, Duc-Tien Dang-Nguyen, Concetto Spampinato, Dag Johansen, Michael Riegler, and others. 2017. A holistic multimedia system for

gastrointestinal tract disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 112–123.

[8] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS)*. ACM, 170–174.

[9] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, and others. 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS)*. ACM, 164–169.

[10] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Carsten Griwodz, Thomas de Lange, Kristin Ranheim Randel, Sigrun Eskeland, Duc-Tien Dang-Nguyen, Olga Ostroukhova, and others. 2017. A comparison of deep learning with global features for gastrointestinal disease detection. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[11] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Thomas De Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico Multimedia Task at MediaEval 2018. In *Working Notes Proceedings of the MediaEval 2018 Workshop*.

[12] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Carsten Griwodz, Thomas Lange, Kristin Ranheim Randel, Sigrun Eskeland, Dang Nguyen, Duc Tien, Mathias Lux, and others. 2017. Multimedia for medicine: the medico Task at mediaEval 2017. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567* (2015).

[14] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012).